Linear classification实验报告

陈家豪 19307130210

一、数据处理

本次实验数据集为 Breast cancer dataset

通过sklearn.datasets获取数据

```
cancer = load_breast_cancer()
cancer_x=cancer.data
cancer_y=cancer.target
```

查看数据大小即部分数据：

```
print("加载完毕，数据大小：")
print(cancer_x.shape)
print(cancer_y.shape)
print("前5个数据：")
for i in range(5):
print(cancer_x[i],cancer_y[i])
```

```
开始加载数据
加载完毕，数据大小：
(569, 30)
(569,)
前5个数据：
[1.799e+01 1.038e+01 1.228e+02 1.001e+03 1.184e-01 2.776e-01 3.001e-01
 1.471e-01 2.419e-01 7.871e-02 1.095e+00 9.053e-01 8.589e+00 1.534e+02
 6.399e-03 4.904e-02 5.373e-02 1.587e-02 3.003e-02 6.193e-03 2.538e+01
 1.733e+01 1.846e+02 2.019e+03 1.622e-01 6.656e-01 7.119e-01 2.654e-01
 4.601e-01 1.189e-01] 0
[2.057e+01 1.777e+01 1.329e+02 1.326e+03 8.474e-02 7.864e-02 8.690e-02
 7.017e-02 1.812e-01 5.667e-02 5.435e-01 7.339e-01 3.398e+00 7.408e+01
 5.225e-03 1.308e-02 1.860e-02 1.340e-02 1.389e-02 3.532e-03 2.499e+01
 2.341e+01 1.588e+02 1.956e+03 1.238e-01 1.866e-01 2.416e-01 1.860e-01
 2.750e-01 8.902e-02] 0
[1.969e+01 2.125e+01 1.300e+02 1.203e+03 1.096e-01 1.599e-01 1.974e-01
 1.279e-01 2.069e-01 5.999e-02 7.456e-01 7.869e-01 4.585e+00 9.403e+01
 6.150e-03 4.006e-02 3.832e-02 2.058e-02 2.250e-02 4.571e-03 2.357e+01
 2.553e+01 1.525e+02 1.709e+03 1.444e-01 4.245e-01 4.504e-01 2.430e-01
```

breast cancer 的影响因素有30个，结果为0,1，表示是否患病

对数据进行拆分, test_size=0.2：

```
x_train,x_test,y_train, y_test=train_test_split(cancer_x,cancer_y,test_size=0.3)
```

二、算法

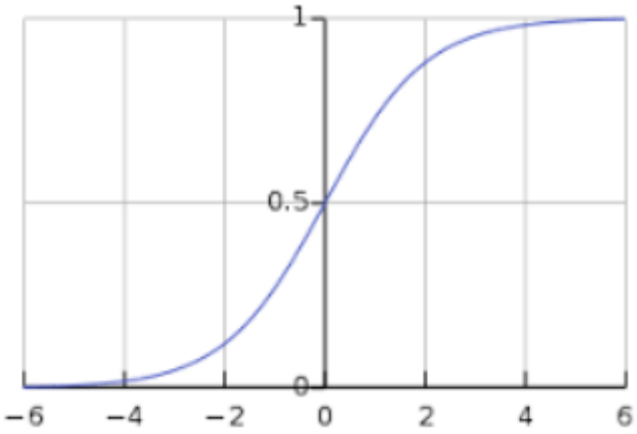
用到 logistic regression 算法

logistic sigmoid function

$$P(\mathcal{C}_1|\Phi) = y(\Phi) = \sigma(\mathbf{w}^\top \Phi)$$

$$\begin{aligned} P(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned}$$

函数图形，结果若大于0.5，归类为1，若小于0.5，归类为0



the likelihood function

$$P(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}, \text{ where } y_n = P(\mathcal{C}_1|\Phi_n)$$

the error function

$$E(\mathbf{w}) = -\ln P(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

which is the cross-entropy error function, where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^\top \Phi_n$

用梯度下降算法求出误差最小的w

三、模型

```
if __name__ == '__main__':

    #开始加载数据
    cancer =load_breast_cancer()
    cancer_x=cancer.data
```

```
    cancer_y=cancer.target
    #分割数据，训练集测试集
    x_train,x_test,y_train,
y_test=train_test_split(cancer_x,cancer_y,test_size=0.2)
    #选择模型
    model = linear_model.LogisticRegression(max_iter=10000)
    #训练模型
    model.fit(x_train,y_train)
    #预测测试集数据的结果
    y_pred = model.predict(x_test)
    #计算判断正确率
    i = 0
    count=0
    for y in y_test:
        if(y==y_pred[i]):
            count+=1
        i+=1
    score = count/i

    print(score)
```

1. 加载数据，并按比例分割成训练集，测试集
2. 选择模型`LogisticRegression`
3. 用训练集数据训练模型
4. 用模型计算测试集数据的结果
5. 计算预测结果与真实结果的正确率

四、结果

```
PS E:\学习\大三（上）\机器学习\作业\Linear_classification> python Linear_classification.py
正确率:
0.9736842105263158
```

模型判断准确率为0.974左右，准确率很高，说明LogisticRegression模型效果很好